Measuring impairment caused by leprosy: inter-tester reliability of the WHO disability grading system

van Brakel, W.H.

published in
Leprosy Review
2004

Lepr Rev (2004) **75,** 221–232

Measuring impairment caused by leprosy: inter-tester reliability of the WHO disability grading system

WILLEMIEN A. NIENHUIS[+],
WIM H. VAN BRAKEL**, C. RUTH BUTLIN* &
TJIP S. VAN DER WERF[+]
**Purulia Leprosy Mission Home & Hospital, West Bengal, India*
***The Leprosy Mission Research Resource Centre, New Delhi, India*
[+]Faculty of Medical Sciences, University of Groningen, The Netherlands

Accepted for publication 1 June 2004

Summary This paper reports the results of a study on the inter-tester reliability of the WHO disability grading system. The WHO disability grading system is the most frequently used method of grading impairment in leprosy patients. With this method, a grade of 0–2 is assigned to each of six individual body sites (both eyes, hands and feet). The maximum grade of any of these sites is used as an overall indicator of the person's impairment status. To date, the WHO disability grading scale has not been subjected to reliability testing. The reliability of the grading system depends on the operational definitions of the grades, the way the tester interprets these definitions and the skill of the tester. It is therefore important that the definitions are unambiguous and leave as little room as possible for multiple interpretations. Three testers with varying degrees of experience did paired assessments on a total of 150 leprosy patients in the Leprosy Mission Hospital Purulia, India, using recently published operational definitions of the WHO disability grades. For every patient, they determined the maximum grade (minimum 0, maximum 2), and calculated the impairment sum-score (EHF score), adding up the six grades for eyes, hands and feet (minimum 0, maximum 12). The weighted Kappa statistic (κ_w) was used as the coefficient of inter-tester reliability. A kappa of 0 represents agreement no better than chance, and 1·0 complete (chance-corrected) agreement. κ_w values of $\geq 0{\cdot}80$ are considered very good and adequate for monitoring and research. Weighted Kappa analysis yielded a reliability coefficient of 0·89 (95%CI 0·84–0·94) for the maximum grade and a κ_w of 0·97 (95%CI 0·96–0·98) for the EHF score. We concluded that, when using standard operational definitions, the WHO disability grading system can be used reliably in the hands of both experienced and inexperienced testers, provided adequate training has been given. Reliability should be evaluated further in a field setting, when used by primary health care workers. It is recommended that the 'WHO disability grading' be renamed 'WHO impairment grading', using the terminology as defined by the International Classification of Functioning, Disability and Health (ICF).

Correspondence to: W. A. Nienhuis (e-mail: WA_Nienhuis@hotmail.com)

0305-7518/04/064053+12 $1.00

Introduction

The World Health Organization (WHO) disability grading system is the most frequently used method of grading impairment in leprosy patients.[1] Since 1960, WHO has advocated a disability classification for use in leprosy.[2] The original 6-point scale has been revised twice. In 1970, the WHO simplified its grading system into a 4-point scale.[3] The proposed classifications were intended to provide a baseline disability status to indicate treatment needs and to monitor changes during follow up.[2,4] The grading system was therefore quite elaborate. However, by 1988 a second revision of the system, a 3-point scale, was proposed.[5] The main purpose of the grading had changed to being a case finding indicator, to estimate delay in case finding. The latest update (1998), in which the grades for the eyes were re-defined, is the grading system used at present (see Appendix 1).[6] With this method, a grade of 0–2 is assigned to each of six individual body sites (both eyes, hands and feet). The maximum grade of any of these sites is used as an overall indicator of the person's impairment status. The validity of the severity weights assigned to different impairments in the WHO grading system has been questioned (Jean Watson, unpublished discussion document on disability grading, 1995).[7]

The WHO disability grading system grades 'impairments' rather than 'disabilities'. We will therefore refer to the 'WHO disability grade' as the 'WHO impairment grade', using the terms as defined in the World Health Organization's International Classification of Functioning, Disability and Health (ICF, formerly ICIDH) (see Appendix 2).[8]

The use of the maximum WHO impairment grade to summarize a person's impairment status has been popularized by the inclusion of this indicator in the International Federation of anti-Leprosy Associations (ILEP) and WHO statistics.[5] The maximum grade given at the time of diagnosis is used as an indication of early case reporting. Unfortunately, assessing the impairment status at the time of release from treatment or even more frequently is not encouraged. Such indicators could provide at least a crude measure of the effectiveness of prevention of impairment (POID) activities.[7]

Information on the impairment status of leprosy patients is being used for (1) decision making and management of (physical) rehabilitation of individual patients, (2) assessing the effectiveness of a leprosy prevention of impairment and disability (POID) programme in preventing the development of (further) impairments and disabilities and treatment of pre-existing ones, and (3) planning of resources needed for treatment and care of people with impairments and disabilities, both before and after completion of drug therapy.[9,10] An impairment classification provides a uniform language for communication between health workers and health centers and for research purposes.[7]

For all grading systems, knowledge of the reliability of the system is required before objective judgements can be made based on the results. To our knowledge, none of the WHO grading scales has been subjected to reliability testing, though an increasing importance is attached to the results of the WHO impairment grading. To date, only one brief report has been published on the reliability of the (E)HF (Eyes, Hands, Feet) score.[11] The EHF score is based on the WHO impairment grading, adding up the six grades for eyes, hands and feet to a sum-score (minimum 0, maximum 12). The precursor to this sum-score, describing the patient's impairment status in more detail, was introduced in 1994 by de Rijk *et al.*[12] Several recent studies have shown the EHF score to be more sensitive to change than the WHO maximum grade for disability and have shown its operational utility in monitoring and evaluation the impairment status of people affected by leprosy.[1,7,13,14] As a minimum, the

EHF score should be calculated at diagnosis and release from treatment. The difference can then be calculated and classified into 'improved', 'same' or 'deteriorated'. From this, useful indicators such as '% of patients with impairment improved' and '% of patients with impairment deteriorated' can be derived.[7] Because of the increasing use of the EHF score, in this study the sum-score was also calculated for every patient. In this way, both the reliability of the WHO impairment grade and that of the EHF score could be investigated.

Reliability of a grading system does not simply depend on the reliabilities of the testing methods or tools used. Reliability also depends on the operational definitions of the grades and the impairments to be graded, the way the tester interprets these definitions and the skill of the tester. This study examined only the inter-tester reliability, using a particular set of definitions and interpretations of terms in the original grading system, which were mainly based on the recently published operational definitions of the WHO disability grades (see Appendix 3).[15] Operational definitions are needed as the WHO grades are interpreted differently in different programmes. The benefit of clear operational definitions would be twofold. First, when programmes or projects follow definitions that are less prone to misunderstanding and different interpretations, the obtained data will be more reliable and thus more reliable comparisons can be made between programmes or cohorts of patients over time. Second, studies that use the EHF score or the Impairment Summary Form (ISF) will achieve more consistent results, since they also depend on the interpretation of definitions of the WHO impairment grades.

Materials and methods

The study was conducted at the 150-bed TLM Leprosy Home and Hospital in Purulia, India. This hospital is a referral hospital serving Purulia District and adjacent districts in West Bengal and Jharkand.

DESIGN

To evaluate the reliability of the WHO disability grading system itself, rather than the agreement between a particular combination of testers, we chose a design with three different tester pairs, who alternated in the sequence in which the tests were performed. This resulted in six combinations, each rating approximately the same number of subjects (range 21–29). Thus each tester appeared as first and second tester an almost equal number of times, virtually excluding the possibility of a systematic bias in terms of first and second tester in this dataset.

TESTING METHODS

In addition to using standard operational definitions, when striving for optimal reliability, one must ensure that the testing methods used are the same for all testers. The WHO grading keys give only a very brief description of what to test, but neither method nor exact area where to test are specified. In this study, sensory testing was done with Semmes–Weinstein monofilaments of, respectively, 2 and 10 g for hands and feet, asking the patient to point to the site where the stimulus was felt, as described by Bell-Krotoski (1990).[16] Ten sites per hand and ten sites per foot were tested. Criteria for sensory impairment were: for the hand, loss of protective sensation (unable to feel the two gram filament)[17,18] on two or more test sites per

nerve. For the foot, loss of protective sensation (unable to feel the 10 g filament)[19,20,21] on two or more test sites was used. A Snellen's E-chart was used for assessing visual acuity. Although not mentioned in the WHO grading keys, muscle weakness also indicates impairment in nerve function. Muscle strength testing for five commonly affected muscles was added, using a manual muscle strength test according to a standard protocol as described by Brandsma in 1981.[22] The criteria for identifying motor impairment was weakness of ≤ 3 on the modified MRC scale (see Appendix 4).[22,23]

STATISTICAL METHODS

The simplest approach to assessing reliability between testers is to look at exact agreement in the ratings. This approach has two weaknesses. Firstly, it does not take into account the size of the disagreements, and, secondly, some agreement between testers would be expected by chance. A more meaningful answer would be obtained by considering the agreement in excess of the amount of agreement that we would expect by chance. The statistic of choice for this situation is the Kappa (κ). This is a measure of the percentage agreement of the testers (the number of tests in which they agree on the results), corrected for the agreement expected by chance alone. The use of weighting ensures that where testers disagree on the results, bigger disagreements have a bigger effect on the Kappa. To assess inter-tester reliability, we used weighted Kappa (κ_w)[24–26] represented with a 95% confidence interval. Kappa ranges from 0 (agreement no better than chance) to 1 (perfect agreement). According to the modified reference values from Altman[26] after Fleiss,[27] a target of at least 0·60 is defined as good agreement, whereas a target of at least 0·80 very good agreement shows (see Appendix 5). The EHF score data were also examined using the 'mean versus difference plot' recommended by Bland and Altman.[28] The difference between the rater scores was examined using a paired *t*-test.

Data were entered in an Epi Info version 6 database, an integrated software program developed by the Centers for Disease Control in the United States and the World Health Organization, and exported to a Microsoft Excel Worksheet for the calculation of κ_w.

SAMPLE SIZE

There are guidelines for interpreting the weighted Kappa which classify the degree of agreement based on bands with a width of 0·2 kappa units. For this study a standard error target such that the 95% confidence interval of the weighted Kappa $< 0{\cdot}2$ Kappa units seemed appropriate. This implies that $1{\cdot}96 \times \text{SE}$ should $< 0{\cdot}1$, i.e. SE (Kappa) $< 0{\cdot}05$. According to Streiner and Norman,[25] such a precision would require $\sim$150 paired assessments, for an estimated Kappa of 0·80. Three testers therefore did paired assessments on a total of 150 leprosy patients. Each pair of testers tested about 50 patients.

PATIENT SELECTION

Both inpatients and outpatients were used for this study. Any registered patient present at TLM Purulia on the survey days was eligible for inclusion in the study, except for patients who had just had surgery. Staff not involved in the testing randomly selected and sent inpatients and outpatients, who were allocated a study number to take part in the study.

During a planned interim analysis after the first 75 patients had been tested, it was found that relatively few patients with leprosy related eye problems had been enrolled. For that

reason, during the second half of the study, selection of patients was biased towards patients with leprosy related eye problems. Staff not involved in the testing did this again.

TESTERS AND TRAINING

Three testers with varying degrees of experience took part in the study. Two of the testers were members of the local staff of TLM Purulia; a physiotherapist with 3 years of experience in impairment grading and an occupational therapy technician with 17 years of experience in impairment grading. The third tester was a medical student visiting TLM Purulia with 2 weeks of training in assessing the WHO impairment grade of leprosy patients (WAN). This training was mainly done by the physiotherapist, who also gave the occupational therapy technician a short refreshment training of the testing methods.

A written protocol describing the standard operating procedure and the testing methods, including the operational definitions of each of the WHO impairment grades and of the impairments itself, was prepared. Prior to the start of the study, a session was held to discuss this protocol. After that, the three testers did pilot testing on several patients using this information to ensure they were confident in the study procedure, the filling in of the record forms and the interpretation of the operational definitions of the WHO impairment grading system.

TESTING

Every week, one pair of testers did the testing, taking care to change the sequence of first and second tester in every other patient. About 25 patients were tested every week. The testing was performed with the tester blinded to previous results and to the results of his fellow tester. Testing for each patient was usually completed within $1\frac{1}{2}$ h. Early in the study, some results showing unusually large and consistent discrepancies were discussed to find out if the cause was a lack of clarity in the protocol. These discrepancies related to the grading of non-leprosy related impairments (one of the testers did not grade non-leprosy related impairments as grade '2'; then it was found that the protocol was not clear enough) and the assessment of visual acuity (in one test room the line at the bottom was not drawn at 6, but at 5 m from the E-chart). They no longer occurred after a few operational adjustments had been made. Patients with such discrepancies in tester results were included in the data analysis.

DATA RECORDING

A custom-made form produced for this study was used for recording the results of the assessments. The testers had to record visual acuity and any other leprosy related problems in the eyes or the face, if present. They had to record muscle-testing scores and fill in the results of sensory testing on a hand and foot chart, marking visible impairments, if present. The individual score for both eyes, both hands and both feet had to be completed. Finally all the information had to be integrated to form the maximum score and the EHF score.

The patient's details (name, age sex, folder number) were recorded separately, as well as on the consent form.

ETHICS

The TLM South Asia Research Committee (TSARC) and the TLM Research Ethics Committee approved the study protocol. Informed consent was obtained from all patients by the first tester using a Bengali consent form.

Table 1. Agreement WHO impairment grade

	Tester 2			
Tester 1	0	1	2	Total
0	33	3		36
1	4	18	7	29
2	1	5	79	85
Total	38	26	86	150

0, 1, 2 = given maximum grade by the tester.

Results

The mean age of the 150 participating patients (111 men, 39 women) was 35 years (SD = 15, range = 8–90). Tables 1 and 2 show the tester-pair agreement of the WHO impairment grade and the EHF score. The percentages of absolute agreement and the weighted Kappa values for inter-tester reliability are presented in Table 3. For the EHF score, the agreement within 1 grade is included. Table 3 gives a subdivision to show reliability for the individual tester-pairs. Interesting for the EHF score is to determine absolute agreement only for the more difficult 'mid-range grades' (when data in which the patient is graded '0' (without impairment) or '12' (maximum impairment) for both testers are omitted). In Table 4 agreement is being shown for the EHF score 'mid-range grades'.

The Bland and Altman plot of the mean EHF score versus the difference between the raters reveals that all but 5 observations (EHF scores differences) lie within ± 2 SD around zero (see Figure 1). In addition, the graph shows no funnel effect, which would be the case if

Table 2. Agreement in EHF score

	Tester 2													
Tester 1	0	1	2	3	4	5	6	7	8	9	10	11	12	Total
0	33	2	1											36
1	3	9	2		1									15
2	2	1	22	3	1	1								30
3			3	5	2	1	1							12
4			5		12									17
5					1	1		2						4
6						4	2							6
7					1	1	2	1						5
8								1	5					6
9										1				1
10									1	1	3			5
11												3		3
12											2	2	6	10
Total	38	12	33	8	18	8	5	4	6	2	5	5	6	150

0–12 = given EHF score by the tester.

Table 3. Inter-tester reliability of the WHO grades, with subdivision for tester pair

Grading system	Testers	Number of patients tested	Absolute agreement (%)	Agreement within 1 grade (%)	Weighted Kappa	95% CI
WHO impairment grade		150	87		0·89	0·84–0·94
	A, B	44	89		0·92	0·86–0·99
	A, C	55	80		0·85	0·77–0·93
	B, C	51	92		0·91	0·80–1·0
EHF score		150	69	87	0·97	0·96–0·98
	A, B	44	73	89	0·98	0·96–0·99
	A, C	55	64	84	0·96	0·93–0·98
	B, C	51	71	88	0·96	0·94–0·98

CI = confidence interval.

Table 4. Absolute agreement in the EHF score 'mid-range grades'

Grading system	Testers	Number of patients tested	Absolute agreement (%)	Agreement within 1 grade (%)
EHF score		111	58	82
	A, B	28	57	82
	A, C	45	56	80
	B, C	38	61	84

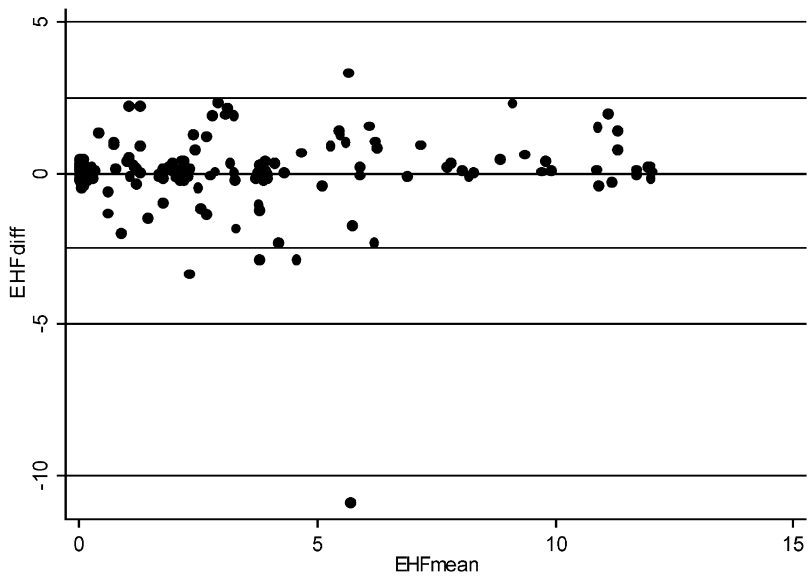

Figure 1. Bland and Altman plot of the difference in between-rater EHF scores versus the mean of these scores ($n = 150$ paired observations).

agreement were proportional to the size of the score. A paired *t*-test of the differences between the raters showed no evidence of any systematic differences in sum score ($P = 0{\cdot}80$).

For the WHO impairment grade, we also examined the weighted Kappa values for the individual grades (data not shown) and found the coefficient for right eye grade to be the lowest (0·85). In itself, this still indicates excellent reliability. Only one paired observation differed by more than one grade. The next lowest was the reliability of the left hand scores (0·88), where 4/150 paired observations differed by two grades and 11/150 by one grade. Such differences do occur occasionally when, e.g. one tester rates a visible scar as '0', while the other rates it as '2', because it constitutes visible impairment.

Discussion

The results of this study show an excellent reliability. Weighted Kappa analysis yielded a reliability coefficient of 0·89 for the WHO impairment grade and a κ_w of 0·97 for the EHF score, both with narrow confidence intervals. Encouraging is the fact that, taking into account the varying degrees of tester-experience, the weakest tester-pair κ_w for the maximum grade was 0·85. The EHF score had a κ_w of 0·96 for the weakest pair. For the maximum grade the percentage of absolute agreement was 87%, while for the EHF score the absolute agreement was somewhat lower (69%). Looking at the absolute agreement within 1 grade, this percentage reached 87%.

It is relatively easy to grade a patient who is 'without impairment' or the opposite, 'maximally impaired'. This results in agreement between testers tending to be best on both extremes of the scales.[29] For that reason, the EHF score absolute agreement was determined in the 'mid-range grades', patients graded '0' or '12' by both testers left out of the analysis. Absolute agreement dropped by 11% to 58%, whereas agreement within one grade reached 82%.

The excellent standard of reliability shown in this study cannot be reasonably expected under most grading circumstances. Our hypothesis is that the use of operational definitions of the WHO disability grading (discussed and practised well before starting the study) is the main source of the current high reliability. Because the WHO grading keys give only very brief guidelines and definitions of the grading system, in practice hospitals or even testers themselves make their own operational definitions. For example, it is unclear where to test and how to define sensory impairment and anesthesia and how to interpret visible impairments (scars, muscle wasting, cracks). Should non-leprosy related impairments be graded (in this study they were included and graded)? Should muscle weakness be included as an impairment (in this study it was)? If yes; how to test and define motor impairment? In case of unclear operational definitions, reliability will probably be less and caution has to be taken to use this 'uniform language for communication' to compare impairment status, e.g. in POID programmes. In addition, it is likely that reliability would be less under field conditions, in contrast to the referral hospital where this study was conducted.[30] In the field, time pressure may adversely affect results, staff motivation may be worse and staff are less experienced and skilled in this type of testing, especially in the near future, when the Government of India is integrating leprosy care into regular government health services. Another difference in the field is the fact that the proportion of patients seen with impairments will be much lower than in cohorts of a referral

hospital. Therefore, reliability should be evaluated further in a field setting, when used by primary health care workers.

Grading of the eyes is difficult and differences in classification and grading results may easily occur. Assessing visual acuity is difficult and not very reliable. In case of impaired vision, is the cause leprosy related? Should cataract be graded as an impairment? In this study, relatively few patients had leprosy related eye problems. MDT has greatly reduced the incidence of eye disease in leprosy,[31] but in areas where the incidence of leprosy related eye problems is higher than in this hospital-based study population, reliability of the WHO grading is likely to be lower.

In contradiction to the way the WHO impairment grading system grades the eye, in this hospital eyes were graded '0' with a vision of 6/18 or better, '1' when vision was 6/24, 6/36 or 6/60 and '2' when vision was worse than 6/60. This way of grading can be defended in that, when vision gradually gets worse, it seems logical that the grade should gradually get higher. Why should the grade suddenly change from 0 to 2 when vision declines from '6/60' to 'worse than 6/60'? This is the way the WHO system defines the eye grades in relation to vision. It would seem more sensible to grade a vision of, e.g. 6/36 or 6/60 as '1'. However, when not using a Snellen's chart but the counting fingers method, vision cannot be determined more specifically than 'better or worse then 6/60'.

As expected and shown, the reliability of the EHF score was better than the reliability of the maximum score. The EHF score is not a perfect impairment grading system; it has some of the same limitations as the maximum grade method, but it is more informative and more sensitive to change than the maximum grade.[9,32] Because of that, we recommend using the EHF score rather than the WHO maximum impairment grade.

We recommend that standard operational definitions be agreed nationally and internationally and be instructed to relevant staff.

To prevent confusion about the terms 'impairment' and disability, we recommend use of the terminology of the International Classification of Functioning, Disability and Health (ICF), a revised version of the ICIDH, recently published by the WHO.[8]

In conclusion, the results of this study suggest that, when using standard operational definitions, the WHO impairment grading system and the EHF score can be used reliably in the hands of both experienced and inexperienced testers, provided adequate training has been given.

Acknowledgements

We are grateful to the patients and staff of The Leprosy Mission Home and Hospital Purulia for their hospitality and their support. We would like to warmly thank the staff of the Physiotherapy Department for their enthusiasm, their patience and their help in translation and the staff of the Medical Records Department for their great assistance. We thank M. Sathish Kumar Paul and Nirendra Acharya for their willingness to give their time and talents assisting in this study. The work at Purulia Leprosy Mission Home and Hospital is dedicated to the service and glory of God.

This research is funded by grants of the Marco Polo Fund, the J. K. de Cock Foundation, the Groninger University Fund and the Netherlands Leprosy Relief.

References

[1] Broekhuis SM, Meima A, Koelewijn LF *et al*. The hand-foot impairment score as a tool for evaluating prevention of disability activities in leprosy: an exploration in patients treated with corticosteroids. *Lepr Rev*, 2000; **71:** 344–354.
[2] WHO Expert Committee on Leprosy, *Second Report*, 189. WHO Technical Report Series, Geneva, 1960.
[3] WHO Expert Committee on Leprosy, *Fourth Report*, 459. WHO Technical Report Series, Geneva, 1970.
[4] WHO Expert Committee on Leprosy, *Fifth Report*, 607. WHO Technical Report Series, Geneva, 1977.
[5] WHO Expert Committee on Leprosy, *Sixth Report*, 768. WHO Technical Report Series, Geneva, 1988.
[6] WHO Expert Committee on Leprosy, *Seventh Report*, 874. WHO Technical Report Series, Geneva 1998.
[7] van Brakel WH, Reed NK, Reed DS. Grading impairment in leprosy. *Lepr Rev*, 1999; **70:** 180–188.
[8] World Health Organization. *International classification of functioning, disability and health (ICF)*. WHO, Geneva, 2001.
[9] van Brakel WH, Anderson AM. Impairment and disability in leprosy: in search of the missing link. *Ind J Lepr*, 1997; **69:** 361–376.
[10] Thappa DM. Disability grading in leprosy: current status. *Ind J Lepr*, 1994; **66:** 299–306.
[11] Saunderson P, Currie H, Byass P, Meima A. Intra- and inter-observer variation in assessing the EHF Score. Abstracts of Congress Papers. *Int J Lepr*, 1998; **66:** 73A–74A, DC63.
[12] de Rijk AJ, Gabre S, Byass P, Berhanu T. Field evaluation of WHO-MDT of fixed duration at ALERT, Ethiopia: the AMFES project—1. MDT course completion case holding and another score for disability grading. *Lepr Rev*, 1994; **65:** 305–319.
[13] Meima A, Saunderson PR, Gebre S *et al*. Dynamics of impairment during and after treatment: the AMFES cohort. *Lepr Rev*, 2001; **72:** 158–170.
[14] Saunderson P, Gebre S, Desta K *et al*. The pattern of leprosy related neuropathy in the AMFES patients in Ethiopia: definitions, incidence, risk factors and outcome. *Lepr Rev*, 2000; **71:** 285–308.
[15] Brandsma WJ, van Brakel WH. WHO disability grading: operational definitions. *Lepr Rev*, 2003; **74:** 366–373.
[16] Bell-Krotoski JA. 'Pocket' filaments and specifications for the Semmes–Weinstein monofilaments. *J Hand Ther*, 1990; **Jan–Mar:** 26–31.
[17] Kuipers M, Schreuders T. The predictive value of sensation testing in the development of neuropathic ulceration on the hands of leprosy patients. *Lepr Rev*, 1994; **65:** 253–261.
[18] Birke JA, Brandsma JW, Schreuders TA, Piefer A. Sensory testing with monofilaments in Hansen's disease and normal control subjects. *Int J Lepr*, 2000; **68:** 291–298.
[19] Hammond CJ, Klenerman P. Protective sensation in the foot in leprosy. *Lepr Rev*, 1988; **59:** 347–354.
[20] Birke JA, Sims DS. Plantar sensory threshold in the ulcerative foot. *Lepr Rev*, 1986; **57:** 261–267.
[21] Rith-Najarian SJ, Stolusky T, Gohdes DM. Identifying diabetic patients at high risk for lower-extremity amputation in a primary health care setting. *Diab Care* 1992; **15:** 1386–1389.
[22] Brandsma W. Basic nerve function assessment in leprosy patients. *Lepr Rev*, 1981; **52:** 161–170.
[23] MRC. *Aids to the investigation of peripheral nerve injuries*, memo no. 7, 2nd edition. HMSO, London, 1962.
[24] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psych Bull*, 1968; **70:** 213–220.
[25] Streiner DL, Norman GR. *Health measurement scales. A practical guide to their development and use*. Oxford University Press, Oxford, 1995.
[26] Altman DG. *Practical statistics for medical research*. Chapman & Hall, London, 1991.
[27] Fleiss JL. *Statistical methods for rates and proportions*, 2nd edition. Wiley, New York, 1981.
[28] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, 1986; **Febr 8:** 307–310.
[29] Brandsma JW, van Brakel WH, Anderson AM, *et al*. Intertester reliability of manual muscle strength testing in leprosy patients. *Lepr Rev*, 1998; **69:** 257–266.
[30] van Brakel WH, Khawas IB, Gurung KS *et al*. Intra- and inter-tester reliability of sensibility testing in leprosy. *Int J Lepr*, 1996; **64:** 287–298.
[31] Waddell KM, Saunderson PR. Is leprosy blindness avoidable? The effect of disease type, duration and treatment on eye damage from leprosy in Uganda. *Br J Ophthalmol*, 1995; **79:** 250–256.
[32] Buddingh H, Idle G. Grading impairment in leprosy [letter]. *Lepr Rev*, 2000; **71:** 85–88.

Appendix 1. WHO disability grading system 1998[6]

Hands and feet	
Grade 0	No anesthesia, no visible deformity or damage
Grade 1	Anesthesia present, but no visible deformity or damage
Grade 2	Visible deformity or damage present
Eyes	
Grade 0	No eye problem due to leprosy; no evidence of visual loss
Grade 1	Eye problems due to leprosy present, but vision not severely affected as a result of these (vision: 6/60 or better; can count fingers at six meters)
Grade 2	Severe visual impairment (vision: worse than 6/60; inability to count fingers at 6 m), also includes lagophthalmos, iridocyclitis and corneal opacities

Appendix 2. Definitions used by ICF[8]

The ICF defines impairment and disability as follows:

Impairment: 'A loss or abnormality in body structure or physiological function (including mental functions). Abnormality here is used strictly to refer to a significant variation from established statistical norms (i.e. as a deviation from a population mean within measured standard norms) and should be used only in this sense'.

Disability: 'An umbrella term for impairments, activity limitations and participation restrictions. It denotes the negative aspects of the interaction between an individual (with a health condition) and that individual's contextual factors (environmental and personal factors).

Appendix 3. Operational definitions and interpretations of the WHO impairment grading system used in this study

a. Hands and feet, visible impairment/deformity (visible impairments should be graded '2') *Absorption (shortening of digits)*: a definite visible impairment, should be graded '2'.

Callous (hard skin that developed because of stress on the skin): normal skin will adapt to stress by hardening. Callous should not be considered an impairment and should not be graded.

Contractures (clawing of digits due to contractures of the joint capsules and/or tendons): should be graded '2'.

Cracks (discontinuity in the skin surface due to dryness of skin, subcutaneous tissue may not be involved): minor cracks should not be graded. Many people develop small cracks, especially during certain times of the year when the weather is cold and/or dry. These occur especially on the side of the heel. Severe cracks should be graded '2'. They may especially occur on the side of the heel and under the first metatarsal head. A severe crack can be defined as a crack where the skin is definitely broken, often with 'raw' tissue exposed at the bottom.

Mobile clawing (clawing of fingers or toes due to muscle weakness/paralysis. The fingers or toes can be passively straightened): should be graded '2'.

Severe muscle atrophy (a sign of severe involvement of the nerve; even in the absence of obvious clawing): should be graded '2'.

Scars (indicative of (a) healed ulcer(s) especially on the plantar surface of the foot): should be considered grade '2' when sensation is impaired, '0' when sensation is normal. (The authors prefer grading scars as '0' when sensation is normal and '1' when sensation is impaired, but we did not want to change too much in the centre's way of grading at once. This would confuse and negetively influence the results of the study.)

Ulcers and wounds (discontinuity of the skin ('broken' skin) involving subcutaneous tissue. Complicated if pus or if rough bone can be felt during gentle probing): should be graded '2'.

b. Eyes

Grade '1': refers to every eye problem due to leprosy except the 'severe ones' (see later), e.g. irregular or absent blink and / or visual impairment of 6/18, 6/24 or 6/36.

Grade '2'(severe eye problems): includes lagophthalmos, iridocyclitis or corneal opacities, and/or visual acuity of less than 6/60.

Impairment of the face: collapse of nose, loss of eyebrows or ear deformities ('severe eye/face problems'). These should be graded '2'.

In general: all impairments that are present should be graded. In case of impairments not related to leprosy they should be recorded by writing the actual cause.

Appendix 4. The modified MRC (Medical Research Council) scale[22,23]

In this study, the following muscles will be tested: abductor digiti minimi, abductor pollicis brevis, wrist extensors, foot dorsiflexors and orbicularis oculi.

Hands and feet	MRC grade	Eyes
Full ROM, full resistance	5	Normal muscle strength
Full ROM, reduced resistance	4	Closes, stays closed against some resistance
Full ROM, no resistance	3	Closes, no resistance (may be gap)
Reduced ROM, some joint movement	2	Gap on strong closure
Flicker only	1	Flicker only
Full paralysis	0	Complete paralysis

ROM = range of movement.

Appendix 5. Reference values for κ (κ_w) modified from Altman[26] after Fleiss[27]

Value of κ (κ_w)	Strength of agreement
$\leq$ 0·40	Poor
0·41–0·60	Moderate
0·61–0·80	Good
0·81–1·00	Very good